УДК 004.048:004.622

Диков А.В., Васяева Т.А.
Донецкий национальный технический университет
кафедра автоматизированных систем управления

ПОСТРОЕНИЕ ДЕРЕВА РЕШЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ОСЛОЖНЕНИЙ НОВОРОЖДЕННЫХ У БОЛЬНЫХ САХАРНЫМ ДИАБЕТОМ

Аннотация

Диков А.В., Васяева Т.А. Построение дерева решения для прогнозирования осложнений новорожденных у больных сахарным диабетом. *Рассмотрена задача прогнозирования осложнений новорожденных у больных сахарным диабетом. Разработано программное обеспечение для построения дерева решений. Выполнено построение дерева решений для прогнозирования осложнений новорожденных у больных сахарным диабетом.*

Ключевые слова: *прогнозирование, дерево решений, беременность, сахарный диабет, осложнения.*

Введение. В настоящее время проблема организации сбора, обработки и анализа информации, полученной в процессе медицинской деятельности, является одной из наиболее актуальных и нерешенных проблем. Медицинские работники в процессе своей деятельности значительную часть рабочего времени посвящают сбору, обработке и анализу медицинской информации, а также диагностике, прогнозированию, выбору оптимального пути лечения или плана профилактических мероприятий. Например, сахарный диабет [1] при беременности является одной из важнейших проблем в современном акушерстве, так как эта патология связана с большим числом акушерских осложнений, высокой перинатальной заболеваемостью и смертностью, и неблагоприятными последствиями, как для ребенка, так и для матери [1-4].

Для обработки и анализа информации в данном случае целесообразно использовать деревья решений, представляющие собой интеллектуальные модели. Такие модели обладают высокими обобщающими способностями и хорошо интерпретируются людьми-специалистами в прикладных областях, которые, как правило, не знакомы с методами и моделями искусственного интеллекта. При решении задачи прогнозирования осложнений новорожденных у больных сахарным диабетом с помощью дерева решений появится возможность осуществлять прогнозирование на этапе беременности, на основе полученных предполагаемых осложнений новорожденного врач сможет производить оптимальный выбор направления лечения или плана профилактических мероприятий для беременной.

Целью работы является построение дерева решений для прогнозирования осложнений у беременных болеющих сахарным диабетом.

Постановка задачи. Для обучения модели мы будем использовать реальные медицинские данные, полученные при обследовании беременных женщин, болеющих сахарным диабетом. Эти данные будут нашим множеством объектов:

$$I = \{i_1 , i_2 , \ldots, i_l, \ldots, i_n\}, \quad (1)$$

где i_j − исследуемый объект, т.е., беременная женщина. Каждый объект характеризуется набором переменных:

$$I_j = \{x_1 , x_2 , \ldots, x_h, \ldots, x_m, y\}, \quad (2)$$

где x_h − независимые переменные, значения которых известны и на основании которых определяется значение зависимой переменной *y*.

Каждая переменная x_h может принимать значения из некоторого множества:

$$C_h = \{c_{h1} , c_{h2}, \ldots\} \quad (3)$$

В данном примере независимыми переменными являются параметры полученные при обследовании беременных женщин, частично представленные в табл. 1.

Таблица 1. Название параметров и диапазон их значений характеризующих беременную

№	Название параметра	Диапазон
1	Возраст	[16;50]
2	Вес	[50;150]
3	Суточная доза инсулина	[30;90]
4	СД болеет с возраста	[1;50]
5	Форма (тяжелая (3), средняя (2), легкая (1))	1,2,3
6	Беременность №,	[1;10]
7	Роды №	[1;10]

Зависимая переменная Y в нашем случае представляет возможные осложнения новорожденного, которые необходимо определять.

Кодирование данных. Кодирование данных нам понадобится для того, чтобы закодировать зависимую переменную Y, так как новорожденный ребенок при рождении может иметь сразу несколько диагнозов осложнений.

Для этого переменную Y представим в двоичном коде, где каждый бит будет соответствовать определенному диагнозу. Перечень всех диагнозов новорожденного и соответствующий ему номер бита представлен в таблице 2.

Таблица 2. Названия диагнозов новорожденного и соответствующий им номер бита

Название диагноза	Номер бита
Пупочная грыжа	1
Бронхопневмания	2
Морфо-функциональная незрелость	3
НГЛД	4
Крупный к сроку гестации	5
Желтуха	6
Перинатальное гипонсически-ишемическое поражение ЦНС	7
Врожденная рассеивание ателектазы легких	8
Д/фетопатия	9
Мертвый ребенок	10
Нет осложнений	11

Для примера, если у ребенка будут диагнозы: д/фетопатия, перинатальное гипонсически-ишемическое поражение ЦНС и бронхопневмания, то переменная Y будет равна 00101000010

Математическая модель. Для построения дерева решений будем использовать метод C4.5 [5-6]. Данный метод является усовершенствованной версией алгоритма ID3. В частности, в новую версию были добавлены отсечение ветвей, а также возможность построения дерева из неполной обучающей выборки, в которой отсутствуют значения некоторых атрибутов.

Общий принцип построения дерева решений, заключается в рекурсивном разбиении множества объектов из обучающей выборки на подмножества, содержащие объекты, относящиеся к одинаковым классам.

Для выбора атрибута разбиения С 4.5 использует критерий, называемый отношением прироста информации. Этот критерий позволяет оценить долю информации, полученной при разбиении, которая является полезной, то есть способствует улучшению классификации:

$$Gain-ratio = Gain(T)/(Split-Info(T)) \quad (4)$$

где

$$Split-Info(T) = -\sum_{i=1}^{n}((|T_i|/|T|)\cdot \log_2(|T_i|/|T|)) \quad (5)$$

Выражение

$$Gain(T) = Info(T) - Infos(T), \quad (6)$$

называется приростом информации. Сама же энтропия узла дерева решений определяется формулой:

$$Info(T) = \sum_{j=1}^{k} p_j \log_2 p_j , \quad (7)$$

и представляет собой сумму всех вероятностей появления примеров, относящихся к определенному классу, умноженную на логарифм этой вероятности.

Энтропия разбиения — это сумма энтропии всех узлов, умноженных на долю записей каждого узла в числе записей исходного множества:

$$Infos(T) = \frac{N_1}{N} \cdot Info(T_1) + \frac{N_2}{N} \cdot Info(T_2) + ... + \frac{N_k}{N} \cdot Info(T_k) \quad (8)$$

.

Выбирается атрибут, который максимизирует выражение (4), т.е. он обеспечит наилучшее разбиение, и будет использован для разбиения в результате чего будут созданы дочерние узлы. К полученным подмножествам применяем такой же подход нахождения оптимального атрибута и продолжаем рекурсивно процесс построения дерева, до тех пор, пока в узле не окажутся примеры из одного класса.

Реализация и результаты. На основе предложенного метода C 4.5 для построения дерева решений было создано программное обеспечение для построения дерева решений по заданной выборке данных.

С помощью разработанного программного обеспечения решалась описанная выше задача по прогнозированию осложнений новорожденных, у беременных болеющих сахарным диабетом. На рис. 1 приведено построенное дерево решений для прогнозирования осложнений новорожденного.

Как видно, в процессе построения дерева были выделены наиболее информативные данные про беременную, болеющую сахарным диабетом, влияющие на диагнозы новорожденного.

Ошибка классификации по построенному дереву решений, вычисленная по данным тестовой выборки, находится в пределах 4-6%.

Вывод. В ходе проведения исследований производился анализ данных беременных болеющих сахарным диабетом с учетом диагнозов их новорожденных. Также был проанализирован алгоритм построения дерева решений методом C 4.5. Разработана программа по построению дерева решений на основе метода C 4.5. С помощью разработанной программы было

построено дерево решений, позволяющее предполагать диагнозы новорожденных, по данным анамнеза беременных болеющих сахарным диабетом.

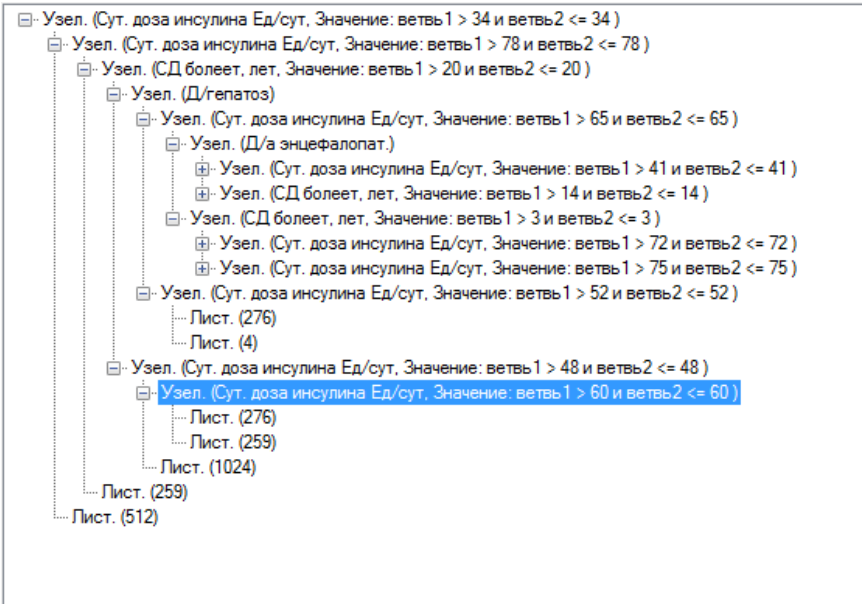

Рисунок 1. Построенное дерево решений

Литература

1. И.М.Грязнова, В.Г.Второва "Сахарный диабет и беременность" Медицина, 1985г.

2. АФП как прогностический показатель состояния новорожденного / Алексеева М.Л., Пустотина О.А., Фанченко Н.Д., Понкратова Т.С. // Проблемы репродукции. — 2005. — № 5. — С. 79-82.

3. Ведмедь А.А. Особенности течения беременности, родов и состояния новорожденных у пациенток с гестационным сахарным диабетом /А.А.Ведмедь, Е.В.Шапошникова"//Вестник Российского университета дружбы народов. 2009.

4. Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиадвигателей : Монография / А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин. – Запорожье : ОАО «Мотор Сич», 2009. – 468с.

5. Classification and regression trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – California : Wadsworth & Brooks, 1984. – 368 p.

6. Rokach L. Data Mining with Decision Trees. Theory and Applications / L. Rokach, O. Maimon. – London : World Scientific Publishing Co, 2008. – 264 p.